

Articulatory Acoustic Feature Applications in Speech Synthesis

Peter Cahill, Daniel Aioanei, Julie Carson-Berndsen

School of Computer Science and Informatics, University College Dublin, Dublin, Ireland

peter.cahill@ucd.ie, daniel.aioanei@ucd.ie, julie.berndsen@ucd.ie

Abstract

The quality of unit selection speech synthesisers depends significantly on the content of the speech database being used. In this paper a technique is introduced that can highlight mispronunciations and abnormal units in the speech synthesis voice database through the use of articulatory acoustic feature extraction to obtain an additional layer of annotation. A set of articulatory acoustic feature classifiers help minimise the selection of inappropriate units in the speech database and are shown to significantly improve the word error rate of a diphone synthesiser.

Index Terms: speech synthesis, unit selection, articulatory acoustic feature extraction

1. Introduction

Unit selection speech synthesisers use a prerecorded speech database as the source of speech units [1, 2]. The output of any unit selection synthesiser will be dependent on the content and the annotations of the speech database. The use of a speech database requires solutions to two significant problems: how to design a synthetic speech corpus that will maximise coverage for the intended domain, and how to avoid errors from the recorded voice. The latter is the focus of this paper.

It is common practice to use phone (or diphone) level annotations of the speech database [3, 4, 5, 6]. The content of the speech database can be annotated manually, automatically or by using a hybrid method of the two. Any of these methods will in practice introduce a degree of error into the system. This is particularly true for systems that use the common forced alignment approach. Forced alignment requires a dictionary file, describing what phones are present and the sequence of them for a given audio stream. Such data in a fully automatic approach are commonly created from the orthographic transcription of the speech database, where a word-to-phones method is used to convert the orthographic transcription into a phone transcription. These orthographic transcriptions are prompted to the reader during the recording of the database.

The problem that arises from this approach is that the reader will sometimes, due to the tedious nature of speech database recording, mis-read what is prompted, without realising it. It is also common for speech synthesis corpora to contain abbreviations or irregular words for which the word-to-phones method will calculate an incorrect pronunciation. Even after these errors are introduced to the system, it is still quite likely that the speaker will not articulate every individual phone perfectly. Any inconsistency between the orthographic transcriptions and the recorded speech will increase the error of the overall speech synthesis system.

While the quality of the forced alignment approach depends on the acoustic model being used, the aligner will typically align all phones of a given sequence listed in the dictionary, even if they are not actually present. Although alignment confidence

values are sometimes available, their accuracy can be misleading. After any speech database is annotated, examination of all instances of a phone will exhibit a significant amount of variance among the different instances of the same phone.

In this paper a technique to obtain an additional, independent layer of annotations for the speech database is presented. For this a set of articulatory acoustic feature classifiers is used that help minimise the selection of inappropriate units in the speech database and thus increase the quality of the synthesised speech.

The remainder of this paper is organised as follows: In Section 2 the use of phonetic annotations in speech synthesis is described. Section 3 describes the synthesis system and explains the relevant articulatory acoustic features. Section 4 defines the testing methodology and the results obtained, and Section 5 concludes and discusses future work.

2. Phonetic Annotations

In order to be able to best identify inappropriate units in the speech database, the concept of having two independent layers of annotations is introduced, a phonological level, and a phonetic level. While phonetic analysis on speech databases for synthesis has been used previously [3], the focus was on acoustic parameters such as MFCCs, F_0 , and power. In this paper this concept is used in a wider sense, focussing on more abstract observations, namely articulatory acoustic features.

The phonological level consists of phone annotations, namely the phone label, diacritics, temporal endpoints and part of speech information. This form of annotation is quite common in modern unit selection systems.

The phonetic level consists of a phonetic analysis of the content of the recorded audio in the speech database. Annotation at the phonetic level is based on discrete time processing of the audio in the speech database and is completely isolated from any other data during the annotation process. This is intentional to avoid the phonological level of annotations from having any influence over the phonetic level.

When the two levels of annotations are obtained, the speech synthesiser can examine both levels looking for consistency between them. This is designed to be an automatic technique where the synthesiser will presume that the most common phonetic properties that exist for any phonological unit being examined will be accepted as the criterion in identifying which units are more suitable for synthesis than others.

The use of two independent levels of annotation results in a cross validation of the annotations. Figure 1 illustrates this, where the region of overlap illustrated between the forced alignment for any phone ($FA(phone)$), and the articulatory features ($AF(f_1, f_2, \dots, f_n)$) indicates the preferred phones.

A situation can arise whereby if there is too much detail in the phonetic level there will be an over-specification of pre-

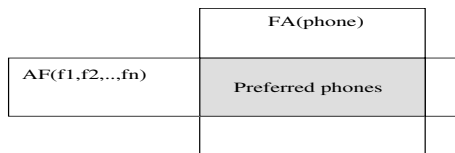


Figure 1: *Phonological (FA) and phonetic (AF) cross validation of the annotations.*

ferred units. Such cases can result in a very small percentage of units being marked as being preferred for synthesis. As a result the unit selection search would omit a significant portion of the speech database. This may be useful in cases where the user seeks a smaller voice database. In the case of this work, however, the goal is to identify the worst units in the speech database which of course will vary significantly between different units. Tests show that typically vowels and noncontinuant units will have a smaller percentage of units preferred for synthesis.

3. Implementation

In the case of the experiments in this paper, the phonetic level of annotation consists of articulatory acoustic features. This involved using multiple independent binary feature extractors that analyse the speech audio and detect the presence of a particular feature.

The system used for testing was a unit selection diphone synthesiser that used a Viterbi search to select the best path through a sequence of diphones, where the join cost is measured by using the Mahalanobis distance measure between the two vectors (\vec{x}, \vec{y}) , such that:

$$D(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})} \quad (1)$$

where Σ is the covariance matrix, and \vec{x} and \vec{y} are the vectors at the diphone joining point of different phones. The vectors contain Mel-frequency cepstral coefficients (MFCCs), F_0 and power. For further information on using this distance measure see [7, 8].

One covariance matrix exists for each phone in the phone set. The observation matrix is populated using a single vector from the diphone joining point of every phone of a given label. All of the math in the distance measure, covariance and observation matrix calculation was done with 64 bit precision.

3.1. Articulatory Acoustic Feature Extraction

For many years, the phone was regarded as the smallest phonological unit that characterised human speech. However, this representation of speech in terms of phones is problematic for some applications and a classification of phones into natural classes is often preferred. These natural classes have defining properties, such as voicing or nasality, which can be described in terms of distinctive features. Jakobson, Fant and Halle [9] and Chomsky and Halle [10] presented the first thorough treatment of phones in terms of bundles of distinctive features. Later Stevens presented a thorough study of the acoustic correlates of these features [11]. A number of studies show that the structure of the vocal tract influences the organisation of such features into feature geometries (or hierarchies) [12, 13]. That is, the way in which features relate to each other is highly influenced by the structure of the vocal tract. Furthermore, Browman and Goldstein [13] model the causal relation between the articulatory gestures and the acoustic properties of features. Thus, artic-

ulator movements produce acoustic/phonetic effects, which in turn can be identified as features. A large number of studies show how these features can be detected from the speech signal. To name but a few, see [14, 15, 16, 17, 18]. The articulatory correlates of the features used in this paper are presented below (descriptions are taken from Halle and Clements [19] and from Chomsky and Halle [10]).

anterior sounds are produced with a primary constriction at or in front of the alveolar ridge; nonanterior sounds are produced with a primary constriction behind the alveolar ridge.

back sounds are produced with the tongue body relatively retracted; nonback or front sounds are produced with the tongue body relatively advanced.

consonantal sounds are produced with a sustained vocal tract constriction at least equal to that required in the production of fricatives; nonconsonantal sounds are produced without such a constriction.

continuant sounds are formed with a vocal tract configuration allowing the airstream to flow through the midsagittal region of the oral tract; noncontinuant or stop sounds are produced with a sustained occlusion in this region.

coronal sounds are produced by raising the tongue blade toward the teeth or the hard palate; noncoronal sounds are produced without such a gesture.

high sounds are produced by raising the body of the tongue toward the palate; nonhigh sounds are produced without such a gesture.

sonorant sounds are produced with a vocal tract configuration sufficiently open that the air pressure inside and outside the mouth is approximately equal; obstruent sounds are produced with a vocal tract constriction sufficient to increase the air pressure inside the mouth significantly over that of the ambient air.

strident sounds are produced with a complex constriction forcing the airstream to strike two surfaces, producing high-intensity fricative noise; nonstrident sounds are produced without such a constriction.

vocalic sounds are produced with an oral cavity in which the most radical constriction does not exceed that found in the high vowels [i] and [u] and with vocal cords that are positioned so as to allow spontaneous voicing; in producing nonvocalic sounds one or both of these conditions are not satisfied.

voiced sounds are produced with a laryngeal configuration permitting periodic vibration of the vocal cords; voiceless sounds lack such periodic vibration.

A set of 19 features were used to train 19 hidden Markov models (HMMs), one for each feature. The choice of which features to use from a set of 19 possible features was based on the reliability of the trained feature extractors. The best 10 (described above) were used in this study. The training technique of the feature extractors used is described in [20]. Each feature extractor was intentionally designed to be independent of the results of any other feature extractors. The next section describes how the features are used by the speech synthesiser.

3.2. Applications

The phonetic level of annotations is used to identify unusual phones. Where for every phone in the phone set, the phonetic features present at the diphone boundary of all phones with a matching label are examined. The most common set of features present at the diphone boundary is noted, and all phones which have the exact most common set of features present at their diphone boundary are identified as preferred phones. Only the diphone boundaries are examined in this process as it is considered to be the most stable point of any phone.

In the tests performed in this work, the preferred phones influence the workings of the synthesiser in two places:

1. Covariance matrix calculation
2. Diphone construction

One covariance matrix (Σ) for the Mahalanobis distance measure is calculated for each phone in the phone set. When calculating the covariance matrices, only diphone boundaries of the preferred phones are used. Phones which are not marked as preferred are ignored as they are likely to distort the variances present between the observations. This is particularly true for cases where the vector parameters present are for a significantly different sound (e.g. frication where a vowel was expected).

When a complete word is unavailable in the database, the system will use diphones to construct the requested word. When the system is constructing diphones for synthesis, the system will attempt to construct diphones using preferred phones only. Only if there is not a continuous pair of preferred phones in the speech database for the required diphone, will the system allow the use of non-preferred phones in the diphone construction.

4. Testing and Results

The voice used for testing was a subset of the ATR Blizzard Challenge 2007 speech data. A subset of the data was intentionally chosen to limit the diphone coverage that existed in the full voice. The aim was to maximise the amount of diphone synthesis to occur during the test, in the case of the full voice data many of the words in the test would have existed in the voice database so the diphone synthesiser would have used continuous units. In order to avoid this situation, a subset of the speech data was used.

4.1. Testing

Three variants of the voice were created for the test. All three were identical with the exception of the phonetic annotations. The differences between the phonetic annotations were:

1. No phonetic level of annotation
2. Preferred phones used
3. Preferred phones used only when the set of preferred phones exceeded 30% of phone coverage

The test utterances were created by making a list of 30 American English utterances. Each utterance was intentionally constructed to use words which are difficult to synthesise and that did not exist in the speech database. All 30 utterances were synthesised using each of the 3 voices. For each participant in the test, 10 of the synthesised utterances from each voice were randomly selected. Participants were supplied with 30 audio files in a random order. Each participant had a different set of audio files and was unaware of which voice was used in each utterance. The participants had to record what was being said in

each utterance. None of the people involved in the creation of the test (including the utterances) were participants for the test.

The participants were prompted with the synthesised utterances without any prior awareness about the content of any of the utterances, their domain or their source. Participants were able to replay an utterance as often as they wished before entering what was being said.

4.2. Results

There were 7 participants in the test, 3 of which were non-Native English speakers but have been resident in an English speaking country for at least 2 years. Results between participants were consistent with the overall mean scores, indicating that the test was sufficient. The word error rate (WER) results are in Table 1.

Voice	WER(%)
1	27.11
2	28.91
3	23.52

Table 1: Results of Word Error Rate tests.

Voice 1 in the test was the synthetic speech system with the use of articulatory acoustic feature extractors disabled. The WER score for Voice 1 was 27.11%, which can be considered a reasonable score for the portion of the speech database used and the intentionally difficult to synthesise test utterances. The results from Voice 1 are similar to results published for other diphone synthesisers using a speech database of similar size with a similar speaker [21]. Voice 2 was the identical synthetic speech system with the feature extractors used for all phones, regardless of the percentage of how many phones were marked as preferred. Interestingly, the WER score is 1.8% higher. This increase in error is due to a situation arising where if the percentage of preferred phones for any phone label is relatively low, too many of the phone candidates are then avoided. The remaining voice, Voice 3 was the voice where the articulatory acoustic feature extractors were only used in cases where it would not involve pruning more than 70% of the available phone candidates for any phone label. This voice showed a significant reduction in error, where the WER score was 23.52%, 5.39% better than Voice 2, and 3.59% better than Voice 1. As expected, this highlights that the results of the synthesiser will be significantly altered by any form of phonetic annotation.

It is anticipated that the ideal percentage threshold at which to decide whether or not to use the feature extractors for any phone will be significantly dependent on the feature set in use, the accuracy of the feature detectors and also the size of the speech database in use. For all 10 features to have the exact same values at the diphone boundaries, 30% seemed to be a reasonable percentage at which to define the percentage threshold. If less feature detectors are used, a higher percentage threshold would perform better. It is likely that larger speech databases can afford for more units to be avoided as there will still be a significant choice remaining. It is also possible that instead of considering the speech database duration as part of the criteria, perhaps the quantity of phones remaining in the set would be a better metric.

5. Conclusions

In this paper, the concept of using both phonological and phonetic annotations of a speech database for synthesis was introduced. Both of these levels of annotation should be obtained independently of each other. The speech synthesiser can then interpret both annotations and make its decision as to what any given unit contains and where it is most suitable for use.

In the case of this work, the phonetic level of annotations consisted of binary articulatory acoustic features. Ten feature extractors were used on the entire speech database to obtain the phonetic level of annotation. The use of the phonetic level of annotation had an influence on the calculation of covariance matrices and on the phone units used during synthesis.

Tests were performed using three variants of a voice on a di-phone synthesiser. The three variants of the voice were derived from the same speech data; one using no phonetic annotations, one using phonetic annotations to avoid phones regardless of the quantity of phones being avoided, and the third used phonetic annotations only when it did not result in avoiding a defined percentage of phones.

The results showed that the overall word error rate of the system significantly improved when using the phonetic annotations in cases where it would not omit any more than 70% of instances of any phone. The degree of error increased when using the phonetic annotations on all phones due to the fact that it avoided too many phones from the database.

Future work will investigate the applications of phonetic annotations in speech synthesis further. Studies using different sets of articulatory acoustic feature extractors are planned, as well as tests on what is the ideal threshold to decide whether or not the phonetic annotations should be used for a particular voice. It is likely that such values will depend on the coverage and the size of the corpus.

6. Acknowledgements

This work was funded by the Irish Research Council for Science, Engineering and Technology (IRCSET), IBM, and SFI. The experiments were done using the ATR Blizzard Challenge 2007 speech database. This material is based upon works supported by the Science Foundation Ireland for the support under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. The authors express their thanks to ATR for access to their speech database. The authors also express thanks to Supphanat Kanokphara for training the feature extractors.

7. References

- [1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP*, vol. 1, 1996.
- [2] A. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," *Proceedings of Eurospeech*, vol. 95, pp. 581–584, 1995.
- [3] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *Proceedings of Eurospeech*, vol. 2, pp. 601–604, 1997.
- [4] M. Beutnagel, A. Conkie, and A. Syrdal, "Diphone synthesis using unit selection," *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves, NSW, Australia, Nov, 1998*.
- [5] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," in *Proceedings of ICASSP*, vol. 1, 2002.
- [6] R. Clark, K. Richmond, and S. King, "Festival 2—build your own general purpose unit selection speech synthesiser," *5th ISCA Speech Synthesis Workshop*, pp. 173–178, 2004.
- [7] A. Gray Jr and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [8] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.
- [9] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press, 1952.
- [10] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, USA: Harper & Row, 1968.
- [11] K. Stevens, *Acoustic Phonetics*. Cambridge, MA, USA: MIT Press, 1999.
- [12] E. Sagey, *The representation of features and relations in non-linear phonology*. PhD thesis, Massachusetts Institute of Technology, 1986.
- [13] C. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology*, vol. 6, pp. 201–251, 1989.
- [14] A. Juneja and C. Espy-Wilson, "An event-based acoustic-phonetic approach to speech segmentation and e-set recognition," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2003.
- [15] S. King, P. Taylor, J. Frankel, and K. Richmond, "Speech recognition via phonetically-featured syllables," *Phonus*, vol. 5, pp. 15–34, 2000.
- [16] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition," *Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1086–1101, 2002.
- [17] S. Chang, S. Greenberg, and M. Wester, "An elitist approach to articulatory-acoustic feature classification," in *Proceedings of Eurospeech*, pp. 1725–1728, 2001.
- [18] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multi-lingual articulatory features," in *Proceedings of ICASSP*, 2003.
- [19] M. Halle and G. Clements, *Problem Book in Phonology: A Workbook for Courses in Introductory Linguistics and Modern Phonology*. MIT Press, 1983.
- [20] S. Kanokphara, J. Macek, and J. Carson-Berndsen, "Comparative Study: HMM & SVM for Automatic Articulatory Feature Extraction," in *Proceedings of the 19th Int'l. Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems*, (Annecy, France), Springer Verlag, June 2006.
- [21] C. Bennett, "Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005," in *Proceedings of Interspeech*, 2005.